Nucleic Acids Research, Vol. 19, No. 16 4333–4339

Evident diversity of codon usage patterns of human genes with respect to chromosome banding patterns and chromosome numbers; relation between nucleotide sequence data and cytogenetic data

Toshimichi Ikemura and Ken-nosuke Wada[1]
DNA Research Center, National Institute of Genetics, Mishima, Shizuoka-ken 411 and [1]Chukyo University, School of Computer Sciences and Cognitive Sciences, Toyota, Aichi-ken 470-03, Japan

Received May 30, 1991; Revised and Accepted July 22, 1991

ABSTRACT

The sequences of the human genome compiled in DNA databases are now about 10 megabase pairs (Mb), and thus the size of the sequences is several times the average size of chromosome bands at high resolution. By surveying this large quantity of data, it may be possible to clarify the global characteristics of the human genome, that is, correlation of gene sequence data (kb-level) to cytogenetic data (Mb-level). By extensively searching the GenBank database, we calculated codon usages in about 2000 human sequences. The highest G+C percentage at the third codon position was 97%, and that of about 250 sequences was 80% or more. The lowest G+C% was 27%, and that in about 150 sequences was 40% or less. A major portion of the GC-rich genes was found to be on special subsets of R-bands (T-bands and/or terminal R-bands). AT-rich genes, however, were mainly on G-bands or non-T-type internal R-bands. Average G+C% at the third position for individual chromosomes differed among chromosomes, and were related to T-band density, quinacrine dullness, and mitotic chiasmata density in the respective chromosomes.

INTRODUCTION

It is increasingly clear that the genome of higher vertebrates is a mosaic of compartments differing in G+C content. By separating genome DNAs by buoyant density fractionation and analyzing the DNAs by Southern blotting and hybridization, Bernardi and his colleagues (1,2, and references therein) found the genome of warm-blooded vertebrates to be a mosaic of very long DNA sequences (>300 kb), each fairly homogeneous in G+C content. They named the mosaic structures 'isochores' and suggested that the structures may be associated with chromosome staining bands, G-(Giemsa dark, Quinacrine bright) or R-(Giemsa pale, Quinacrine dull) bands. Studies on the nucleotide sequence level have also demonstrated the presence of G+C% mosaic structures in higher vertebrate genomes. While studying codon usage in higher vertebrate genes, our group (3–5) and others (6–8) found genes with a high G+C% at the third codon position (e.g. >80% G+C) to almost always be surrounded by long GC-rich (e.g. 55–65% G+C) genome sequences, while those with a low G+C% at the third position (e.g. <50% G+C), by long AT-rich sequences (e.g. 35–45% G+C).

Analyses on G+C% distribution along DNA sequences arranged according to their genetic positions (5,9) also show the genomes of higher vertebrates to be composed of G+C% mosaic structures and the sizes of the mosaic units to far beyond those of genes (presumably larger than several hundred kb). To examine the correlation of this mosaic to chromosomal bands, we classified GenBank Sequences into G- and R-band sequences based on the Human Gene Map, and analyzed G+C% levels of the classified sequences (4,5). A major portion of the G-band sequences was less than 50% G+C, and a major portion of R-band sequences was more than 50% G+C: these are the G+C% levels of the sequences themselves but not those at the third codon position. These findings are roughly consistent with expectations from cytogenetic studies, and discrepancies are largely confined to R-band sequences.

Gardiner, Aissani, and Bernardi (10) recently constructed a compositional map of human chromosome 21 and reported G-bands to correspond to fairly homogeneously AT-rich sequences, but R-bands to be much more heterologous and divisible into subgroups: the terminal R-band of chromosome 21 was GC-rich but the internal R-bands were at lower G+C% levels than the terminal band. Holmquist (11) summarized a wide range of cytogenetic observations including human T-banding patterns (Fig. 1) and discussed in detail biological significance of chromosomal bands and base compositional compartmentalization. T-bands have been shown by microscopic studies to be a subset of R-bands and thought to correspond to an evidently GC-rich genome portion because they represent chromosomal segments most resistant to heat treatment (11,12): they are called T-bands since many of them correspond to terminal R-bands (first band from telomeres). It should be noted that all terminal bands of human chromosomes at 850-band level, except for the short arm of chromosome 3 and the long arm of chromosomes 4 and 19, are R-bands (13). Very GC-rich bands that were identified by contrast subtraction of DAPI (AT-specific) and chromomycin A3 (GC-specific) stained chromosomes are known to mostly

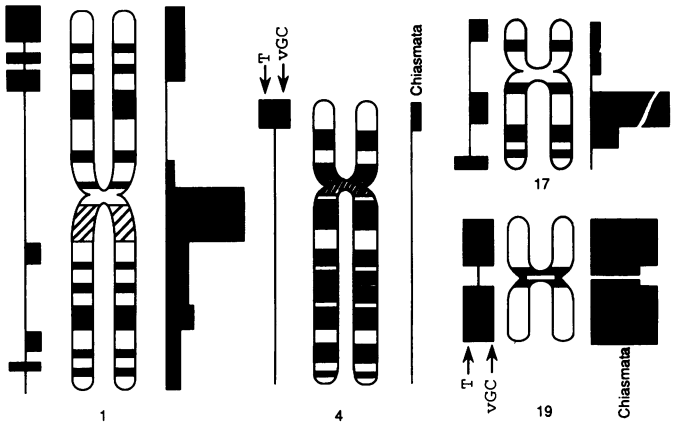

Fig. 1. Schematic representation of chromosomal banding patterns (G, R) with T-banding patterns and chiasmata density distribution. Examples for four human chromosomes are listed; diagrams and banding patters at 300 band-level were redrawn according to Holmquist (11). Very GC-rich bands (v-GC) correspond to bands visualized by double staining with DAPI and chromomycin A3 (23). In the 300-band diagram, the entire portion of chromosome 19 except for the cetromeric portion was assigned to R-band (24).

correspond to T-bands (11; see Fig. 1). In this work, the chromosomal locations of genes evidently GC-rich at the third codon position were found to be mainly on T-bands and/or terminal R-bands.

RESULTS

Evidently GC-rich or AT-rich genes at the third codon position that have been precisely mapped on human chromosomes

Based on a survey of the GenBank database, codon usages of all available gene sequences were compiled and published in the Sequence Supplement of this journal (14). When the primate file of GenBank (Release 65, 1990) was analyzed, the codon usages of 1985 human sequences were calculated. G+C% at the third codon position in 244 human sequences was 80% or more (the highest G+C% was 97%), and that in 148 sequences, less than 40% (the lowest was 27%). The overall G+C% of the human genome is about 40%. Thus, there should be a causative factor to produce GC-rich patterns such as those with 80% G+C or more. The chromosomal locations of the evidently GC-rich (as well as the AT-rich) genes were searched as follows, for possible connections with chromosome banding patterns. Surveying the recent Human Gene Map (HGM 10.5, 1990)(15), 45 out of the 244 GC-rich genes with more than 80% G+C at the third position, and 22 out of the 148 AT-rich genes with less than 40% G+C, were found to be on a unique R- or G-band at either 400- or 850-band resolution. Table 1 lists chromosomal positions and band characteristics of these precisely mapped genes (the GC-rich genes in the upper part and AT-rich genes in the lower part). Interestingly, only four GC-rich genes were on G-bands, and the remaining 41 on R-bands. Furthermore, 34 of the 41 R-band genes belong to T-bands and/or terminal R-bands. Thus the evidently G+C-rich genes were mainly on specific genome portions.

The chromosomal positions of the AT-rich genes were clearly distinct from the GC-rich genes. Out of 22 AT-rich genes, 13 were on G-bands and the remaining 9 on R bands. Four out of 9 R-band genes (and thus, out of the total 22 AT-rich genes) were on T-type R-bands, and no AT-rich genes were on terminal R-bands. This contrasts with the finding that GC-rich genes were mainly on terminal and/or T-type R-bands. It should be noted also that many GC-rich genes, but essentially no AT-rich genes,

Table 1. Evidently GC-rich or AT-rich genes at the third codon position. Seq.; GenBank LOCUS name (see also 14). Map; chromosomal band position. GC%; G+C% at the third codon position. Gene; gene symbol. G/R; G- or R-band, the band at the 400-band level is specified by a dash ('). Ter; terminal band. T; T-band, that were usually assigned at the 300- or 400-band level. Chia.; mitotic chiasmata density was according to Holmquist (11), evidently high (+). GC-rich and AT-rich genes are at the upper and lower parts, respectively, and genes in each part are according to chromosome number. When redundant sequences for one gene, including those due to alternate splicing, were available, the longest one was selected in this and the following analyses.

(GC - rich Sequence)

Seq.	Map	GC%	Gene	G/R	Ter	T	Chia.
HUMPOMC	2p23	89.18	POMC	R			
HUMMHCD8A	2p12	83.90	CD8A	G			
HUMDES	2q35	81.91	DES	R			
HUMALPPB	2q37	80.04	ALPP	R'	Ter	T	
HUMCP21OHC	6p21.3	80.65	CYP21	R		T	+
HUMMHCA1A	6p21.3	80.33	HLA-A	R		T	+
HUMMHCACA#1	6p21.3	81.35	HLA-C	R		T	+
HUMHSP70D	6p21.3	92.04	HSPA1	R		T	+
HUMNCF1A	7q11.23	81.59	NCF1	R		T	
HUMCRF	8q13	81.73	CRH	R			
HUMMYCPOB	8q24	86.43	MYC	R'	Ter	T	
HUMDBHRA	9q34	82.28	DBH	R'	Ter	T	+
HUMCATDC	11p15.5	86.44	CTSD	R	Ter	T	
HUMRASH	11p15.5	81.05	HRAS	R	Ter	T	
HUMGFI2	11p15.5	80.66	IGF2	R	Ter	T	
HUMINS01	11p15.5	80.18	INS	R	Ter	T	
HUMHTH1R	11p15.5	86.14	TH	R	Ter	T	
HUMINT2	11q13	83.75	INT2	R'		T	+
HUMHST	11q13.3	93.72	HSTF1	R		T	+
HUMINT1G	12q13	83.56	INT1	R'		T	+
HUMCFVII	13q34	80.51	F7	R	Ter		
HUMFX8*	13q34	80.98	F10	R	Ter		
HUMCKBBA	14q32.3	90.31	CKB	R	Ter	T	
HUMIGHAE2*	14q32.3	80.17	IGHE	R	Ter	T	
HUMHBA4#1	16p13.3	88.81	HBA1	R	Ter	T	
HUMHBA4#2	16p13.3	88.81	HBA2	R	Ter	T	
HUMXIHB	16p13.3	94.41	HBZ	R	Ter	T	
HUMHBQ1A	16p13.3	90.91	HBQ1	R	Ter	T	
HUMMT2A	16q13	80.65	MT1G	R			
HUMAPRTA	16q24	81.22	APRT	R	Ter	T	+
HUMAICEB	17q23	81.33	DCP	R			+
HUMGAAA	17q23	84.78	GAA	R			+
HUMBCL2B	18q21.3	83.50	BCL2	R			
HUMMAG	19q13.1	85.33	MAG	R		T	+
HUMTGFB	19q13.1	83.16	TGFB1	R		T	+
HUMAPOE4	19q13.2	89.62	APOE	G		T	+
HUMJUNCAA	19q13.2	82.18	JUNB	G		T	+
HUMCGB	19q13.3	82.53	CGB	R		T	+
HUMLHB	19q13.3	82.39	LHB	R		T	+
HUMPKCGA	19q13.4	85.32	PRKCG	G	Ter	T	+
HUMPFKLA	21q22.3	83.35	PFKL	R	Ter	T	+
HUMBCR	22q11	82.98	BCR	R'		T	+
HUMG6PDR	Xq28	85.00	G6PD	R	Ter		
HUMUBILP	Xq28	81.65		R	Ter		
HUMCPGISL	Xq28	90.41		R	Ter		

(AT - rich Sequence)

Seq.	Map	GC%	Gene	G/R	Ter	T	Chia.
HUMBLYM1	1p32	38.98	BLYM	R'		T	
HUMACADM	1p31	27.25	ACADM	G'			
HUMAMY110*	1p21	34.25	AMY1	G			
HUMAMY210*	1p21	34.18	AMY2	G			
HUMCA1XIA	1p21	27.87	COL11A1	G			
HUMANTLF3	1p13	36.25	CD58	R'			
HUMDAFA#1	1q32	39.79	DAF	R'		T	
HUMMCP	1q32	31.95	MCP	R'		T	
HUMFBRG#1	4q28	39.73	FGG	G			
HUMGAPA	5q13	36.83	RASA	R'			
HUMMCM	6p21	32.57	MUT	R'		T	+
HUMFNRB	10p11.2	38.30	FNRB	R			
HUMCDC2	10q21.1	34.90	CDC2	G			
HUMPRPC	12p13.2	32.83	PRB1	G			
HUMPRPF	12p13.2	37.10	PRB4	G			
HUMPRPH1	12p13.2	38.92	PRH1	G			
HUMPRPH2	12p13.2	39.52	PRH2	G			
HUMRASK25*	12p12.1	33.16	KRAS2	G			
HUMRB1RA	13q14.2	33.69	RB1	G			
HUMCYES1	18q21.3	29.23	YES1	R			
HUMHPRTB	Xq26	39.73	HPRT	R			
HUMZFY	Yp11.3	37.66	ZFY	G	Ter		

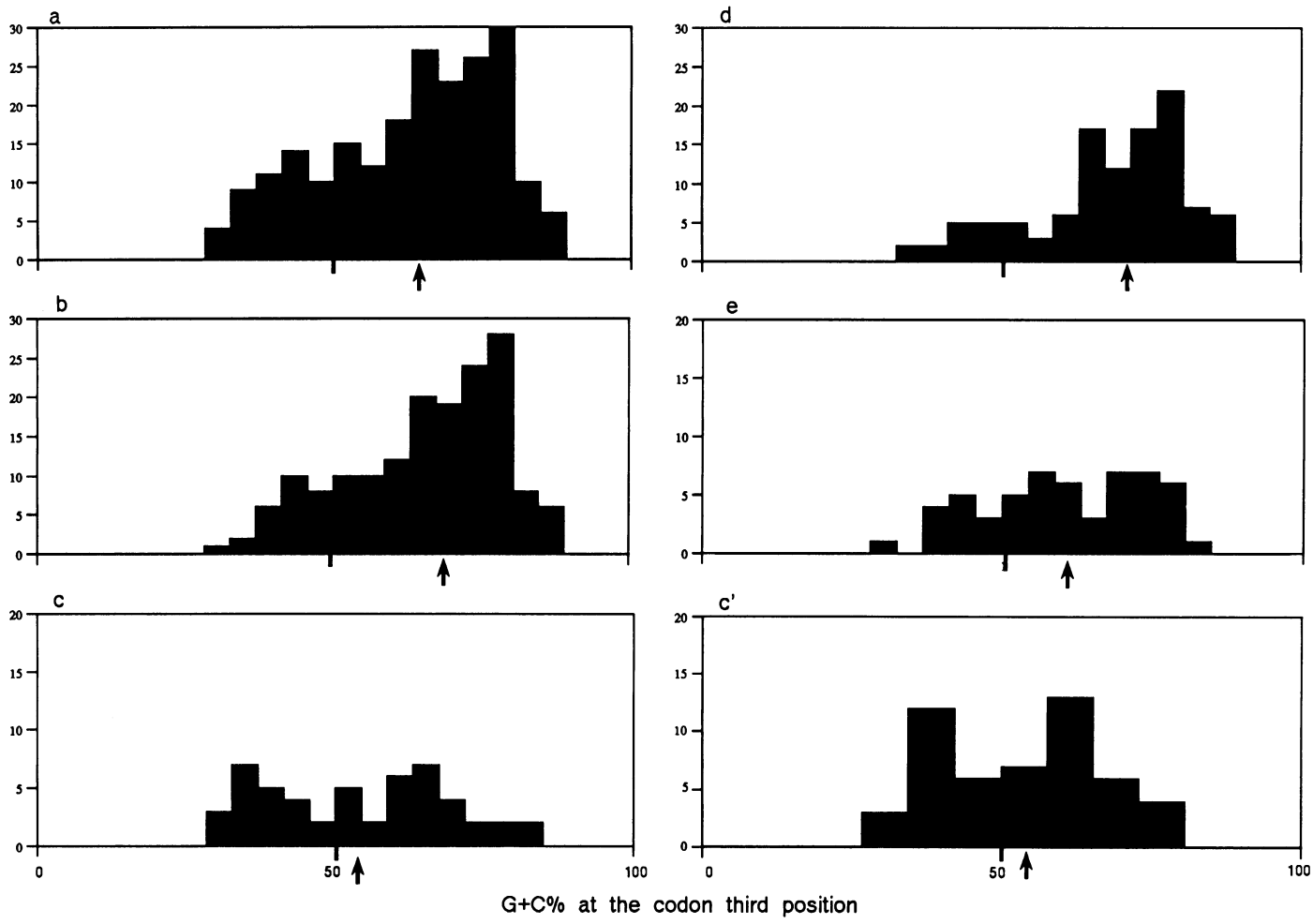

Fig. 2. Histograms for the third codon G+C% of human genes mapped either on a unique R- or G-band. a) R plus G genes, b) R-genes, c) G-genes, d) T-type and/or terminal R-genes, e) non-T-type internal R-genes. A 5% abscissa interval is used for a–e. c') G-genes; because of a small sample size of G-genes compared with others, frequency distribution of a 10% interval is also presented. Average G+C% for each group is indicated by an arrow at the bottom of the graph.

were on mitotic chiasmata dense regions (marked by + in Table 1) that are known to be very GC-rich at a microscope level (11; refer to later discussion). These differential locations of GC-rich and AT-rich genes are consistent with the cytogenetic findings noted by Holmquist (11) and the isochore map constructed for chromosome 21 (10): similar findings were noted by Bernardi and his colleagues (personal communication from G.Bernardi). Cases of minor inconsistencies will be discussed later.

Classification of 600 human genes according to band-types

Codon usages of individual genes should be constrained by a complex combination of various factors (3) and affected by random drift of neutral mutations (16–18). The strategy in Table 1 focuses attention on examples of extreme characteristics such as the evidently GC-rich or AT-rich genes at the third position, and the differential locations of these genes were found. When codon usages of moderate G+C% levels are considered, the effects of factors other than base composition will become significant. It should also be noted that data on chromosome banding are still preliminary. Many narrow or faint bands detectable only at high resolution (e.g. 2000-band level; 19) have not yet been placed in the currently available Human Gene Map. Therefore, a single band even at the 850-band level (see examples in Fig. 4) only means that a major portion of it has the characteristics of the band but some minor portions will correspond to bands with opposite characteristics. [This may explain in part the minor inconsistency between sequence and cytogenetic data in Table 1.] Furthermore there should be certain error in gene mapping in currently available maps. A useful strategy for overcoming these limitations and complications would be to analyze a large quantity of data.

When the Human Gene Map (15) was surveyed, about 600 genes were found to be mapped on a unique R- or G-band either at the 400- or 850-band level (13). Sequences for these precisely mapped genes were sought, surveying the GenBank human sequences. A total of 215 nonredundant gene sequences (164 R-band and 51 G-band genes), for which codon usages have been calculated (14), was selected; the number of R-band genes was higher than that of G-band genes, supporting the notion of the relative paucity of functional genes in G-bands (20). Instead of listing individual data, histograms of G+C% at the third codon position are shown in Fig. 2; (a) all genes, (b) R-band genes (abbreviated to R-genes), (c) G-band genes (G-genes), (c') G-genes with a different abscissa interval from c (see legend), (d) terminal and/or T-type R-genes, (e) internal non-T-type R-genes; non-T-type R-genes mean those on R-bands which are not T-bands. The average G+C% in each group is indicated by an arrow placed

at the bottom of the graph. The distributions of R-genes and G-genes clearly differed (compare b and c; the averages of the R-genes and of the G-genes were 67.7 and 54.9%, respectively). This is consistent with our previous finding in a small scale analysis of genome fragment G+C% (4,5). Distinction was more evident between G-genes (c,c'; 54.9%) and terminal and/or T-type R-genes (d; 70.6%). The distribution of internal non-T-type R genes (e; 61.9%) was intermediate between the two.

Next to be discussed are two major peaks found for G-genes: one peak was 30−40% G+C, and the other, 60−70% G+C (Fig. 2c and c'). G+C% of the second peak was clearly less than the main peak of R-genes (75−85%), and therefore this second peak may not correspond to mis-assigned genes on currently available gene maps. It was previously found that local G+C% levels of protein coding genes, especially those of the genes embedded in AT-rich genome compartments, could be divided into at least two groups (21). In one group, local G+C% levels, as well as those at the third position, were roughly equivalent to those of compartments. In the second group, local G+C% levels and those at the third position were roughly 10% and 20% higher than that of the compartment, respectively. The two major peaks for the G-genes may correspond to these two groups previously found. The biological meaning of the grouping is not known.

Classification of human genes according to chromosome numbers

In the above analyses, attention was directed to genes mapped on a unique R- or G-band at either the 400- or 850-band level. Such genes that have been mapped on an interval ranging over plural bands at the 850-band level (e.g. 1p13.1-p13.2) should thus be excluded from analysis even though their locations are more accurate than those of genes mapped on a unique band at the 400-band level (e.g. 1p13). If such genes are included in the analysis, their affiliation to the G- or R-band must be assessed by ourselves, based on diagrams of the 850-band level. This assessment is apparent in some cases, but rather subtle in others. To depend only on the assignment of the Committee on Human Cytogenetic Nomenclature (13) without our own judgment, only genes mapped on a unique band either at the 400- and 850-band level were selected. Genes thus excluded in Fig. 2 and those much less precisely mapped were analyzed in a separate way.

Cytogenetic data on several aspects indicated G+C% levels to significantly differ for different chromosomes. Quinacrine brightness is known to be positively related to the AT-richness of genome DNA (and thus inversely related to GC-richness)(11,20). Kuhn (22) extensively analyzed quinacrine brightness for each human chromosome, and found clear differences in average brightness among different chromosomes (Table 2). The highest brightness (and thus highest AT-richness) was found for chromosome 4 and the lowest brightness (the highest GC-richness), for chromosomes 19 and 22; the brightness of the two groups differed by a factor of 4. T-band density in individual chromosomes also differs among chromosomes (Fig. 1). The T-band fraction is very high in chromosomes 19 and 22, but very low in chromosome 4.

By analyzing genes mapped on individual chromosomes (thus including genes not yet precisely mapped), we determined whether codon usages differ among chromosomes or not in the following ways. First, human genes for which codon usage could be calculated were classified according to chromosome number mainly based on comments in the FEATURE section of GenBank entries. Chromosomes for which at least 30 nonredundant genes were available were selected. Ten chromosomes and about 390 nonredundant gene sequences belonging to the ten chromosomes were thus selected. The map locations of these genes were then checked using HGM 10.5, because in the FEATURES of GenBank entries, information of HGM 10.5 has not been fully included and clearly there is error in gene assignments and map locations. After this verification, 374 genes belonging to the ten chromosomes were selected, and the average G+C% at the third codon position for each chromosome was calculated (Table 2). The average was from 75% to 49%, and the highest G+C% was that of the evidently T-band rich and quinacrine dull chromosome (Chr. 19) while the lowest, that of the evidently T-band poor and quinacrine bright chromosome (Chr. 4).

Table 2. Average G+C% at the third codon position, T-band fraction, quinacrine brightness value, and chiasmata density for individual chromosomes. Chr.; chromosome number. No.Seq.; number of sequences analyzed. GC%; average G+C% at the third codon position for each chromosome. T-frac.; T-band proportion calculated from the diagram of Holmquist (11). Qbright. and Chi.Den.; average quinacrine brightness and chiasmata density for each chromosome were from Kuhn (22). Data for sex chromosomes were not presented by them.

Chr.	No.Seq.	GC%	T-frac.	Qbright.	Chi.Den.
19	36	75.00	0.81	0.20	5.62
17	36	74.60	0.42	0.35	3.18
6	36	65.80	0.11	0.66	1.87
11	44	65.80	0.22	0.55	1.31
X	32	60.10			
7	35	60.00	0.23	0.61	0.44
12	34	58.40	0.21	0.49	1.05
5	30	58.00	0.14	0.57	0.59
1	59	56.20	0.25	0.56	0.98
4	32	49.00	0.08	0.79	0.03

T-band proportion, quinacrine dullness, chiasmata density, and codon GC-richness for individual chromosomes

To obtain much general data on the correlation between the T-band proportion and GC-richness of the codons, we estimated the T-band proportion in each chromosome simply by measuring the T-band proportion in the schematic representation drawn by Holmquist (11; see Fig. 1). The result is shown in Table 2. Actual T-banding patterns (12) are undoubtedly more complicated than the diagram drawn by Holmquist, and thus the above estimation is rather rough. However, the diagram appears to represent well the global characteristics of individual chromosomes since the diagram for very GC-rich bands visualized by a totally different staining technique, i.e. double staining with DAPI and chromomycin A3 (23), gave essentially the same pattern (Fig. 1). Furthermore the T-band fraction thus estimated and quinacrine brightness values (indices for AT-richness) described by Kuhn (22) for individual chromosomes gave a clear negative correlation (Fig 3a; correlation coefficient $r = -0.92$), indicating the former value to represent well the base compositional levels of chromosomes.

When the T-band proportion and average codon G+C% of the respective chromosome were plotted, a positive correlation was obtained (Fig. 3b; $r=0.75$). This shows the tendency of the genes with GC-rich codons to be enriched in T-band rich chromosomes. Figure 3c examines the correlation between quinacrine brightness and codon G+C% for individual chromosomes. A negative correlation ($r=-0.82$) was found, again showing the codon G+C% to reflect the general characteristics of the base composition in individual

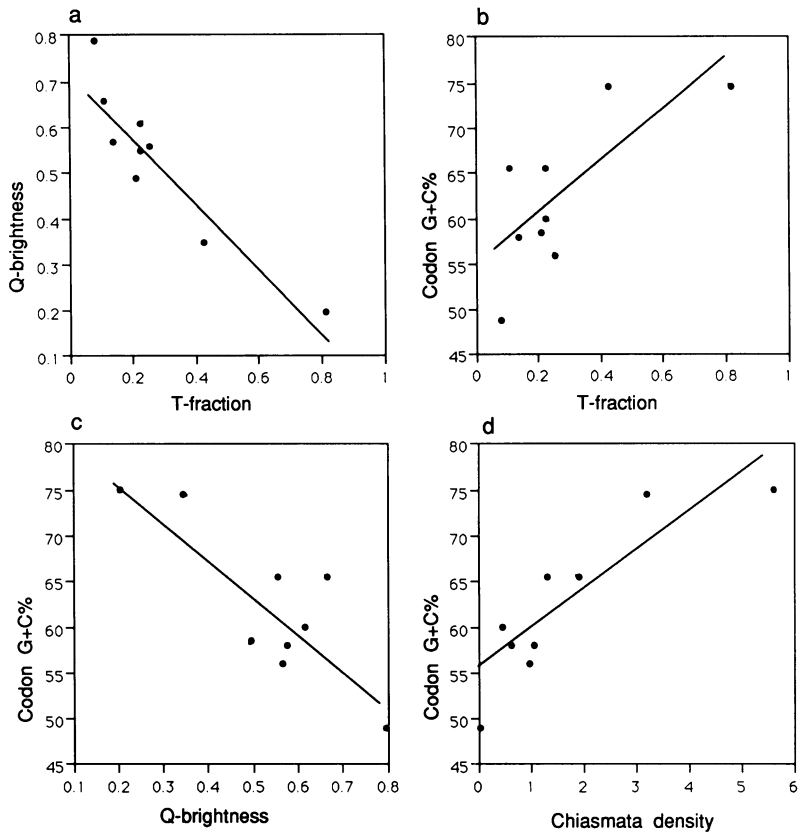

Fig. 3. Relationship of average codon G+C% to cytogenetic data for individual chromosomes. A straight-line that fits data points basing on least squares regression is presented in each graph. a) T-band fraction and quinacrine brightness; correlation coefficient r=−0.92. b) T-band fraction and G+C% at the third position; r=0.75. c) Quinacrine brightness and G+C% at the third position; r=−0.82. d) Chiasmata density and G+C% at the third position; r=0.87.

chromosomes. The frequency of Bloom syndrome mitotic chiasmata is known to be related to GC-richness of the genome portion (22) and is high in T-bands (11); Bloom syndrome is a heritable human disease caused by a defective DNA ligase I enzyme (25). Actually, the locations of many evidently GC-rich genes in Table 1 corresponded to the chiasmata-dense region (marked by +). Kuhn (22) extensively analyzed the chiasmata density of each chromosome (number of chiasmata found for each chromosome divided by length of the chromosome), and found it to differ significantly among chromosomes and related to the GC-richness of the chromosomes. This density (Table 2) and the codon G+C% were plotted (Fig. 3d) and showed a positive correlation (r=0.87). Results in Table 2 and Fig. 3 thus show that the sequence data and the cytogenetic data obtained on several different aspects are closely related, and codon usage is a sensitive tool to understand the global characteristics of chromosomes.

Codon G+C% levels of genes on GC-rich or AT-rich chromosomes

To explain that differential codon G+C% levels among chromosomes (Table 2) are not mere reflection of characteristics of genes localized in a restricted area of the chromosomes, G+C% levels of all genes were listed by vertical bars for the AT-richest chromosome (Chr. 4; Fig. 4a), and GC-richest (Chr. 19; 4b) and second GC-richest (Chr. 17; 4c) chromosomes. Short thicker bars were used for accurately mapped genes (i.e. within a few bands), and long thinner bars for those less accurately mapped; several long thicker bars corresponded to multiple thin bars that happened to be near each other. Clearly, most of the genes of chromosomes 17 and 19 are more GC-rich than those of chromosome 4. This is consistent with the cytogenetic finding that a large portion of the former chromosomes corresponds to T-bands. In this connection, a few discrepancies between sequence data and cytogenetic data in Table 1 (i.e. four GC-rich sequences belonged to G-bands) will be discussed. Three of the four were found to be on the G-bands in a broad T-band of the GC-richest chromosome (Chr. 19; Fig. 1).

DISCUSSION

Different types of cytogenetic data

The strategy in this work was to survey many codon usage patterns and summarize them. The global characteristics of the genome, often unnoticed in small scale analyses, were found, and their relationships to several types of cytogenetic data (G-band, R-band, T-band, quinacrine brightness, and chiasmata density) thought related to base composition, were studied. Actual banding patterns of chromosomes are known to be more complicated than observed in the black (G-band) and white (R-band) diagram, and bands within each class to vary in intensity (13). [The diversity found for each band class in Fig. 2 may be partly due to these factors.] Therefore it is usually difficult to accurately quantify the cytogenetic data, and quantitative estimation made for individual data is rather rough. Furthermore,

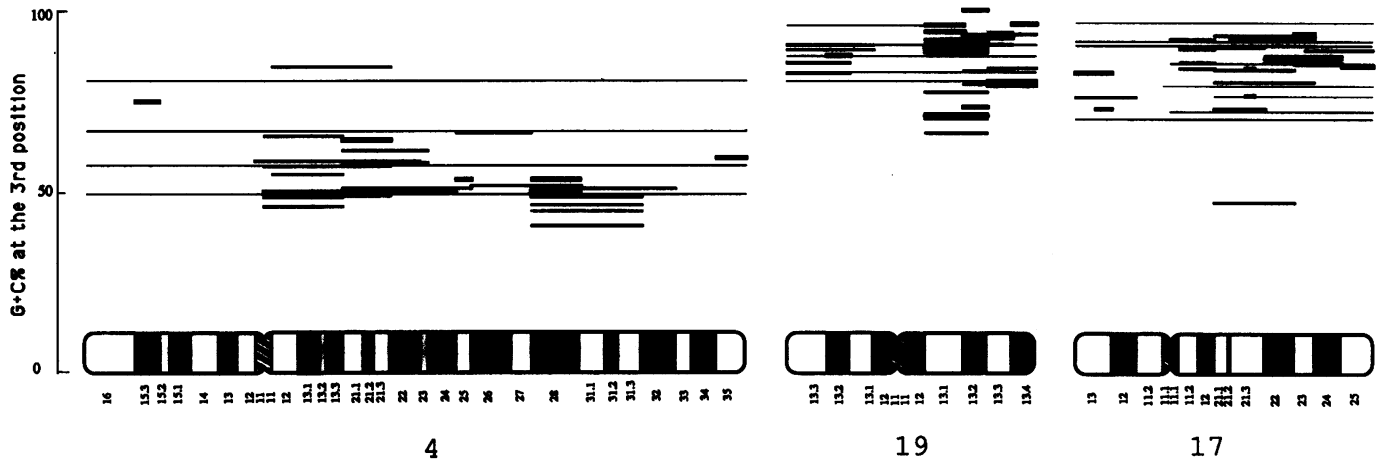

Fig. 4. Codon G+C% of genes mapped on the chromosome 4, 17 or 19. Height of the horizontal line ranging over a mapped area shows G+C% at the third codon position of the gene. Chromosome band patterns at 850-band level are listed.

precise mechanisms of chromosome banding are still unclear although G+C% compartmentalization is important and possibly a major determinant. Factors other than compositional distribution must be involved in certain banding. For instance, chromatin structures appear to be involved in G-banding while the chromatin structures themselves most likely are related to base composition (11,20,26). Because of this uncertainty, different types of cytogenetic data obtained by different techniques have been used here even though possibly somewhat redundant. At first glance, the conclusion that human codon usage patterns are related to cytogenetic data appears rather odd because the levels of the two phenomena are so different. It is increasingly clear, however, that codon usages are closely related to isochore structures (hundred kb-level phenomenon) and isochore structures closely related to chromosome banding structures (2). Thus, this conclusion is consistent with accumulating evidence.

T-bands and terminal R-bands

The rather vague expression, 'T-bands and/or terminal R-bands' is used in this paper. Many T-bands correspond to terminal R-bands, but there are internal T-bands, as well as non-T-type terminal R-bands. So far, concerning the 215 genes in Fig. 2, the average G+C% for each of the following 6 cases is as follows: T-type terminal R-bands (35 such sequences, average, 75.7%), terminal R-bands (54 Seq., 73.4%), T-type R-bands (90 Seq., 71.5%), T-type internal R-bands (55 Seq., 67.9%), non-T-type terminal R-bands (19 Seq., 66.7%), and non-T-type internal R-bands (55 Seq., 64.3%). Two characteristics, T-band and terminal positioning, are related to increase in G+C%. [There are a few T-type G-bands, that are rather exceptional and correspond to G-bands at the 850-band level located within broad T-bands at the 400-band level. The exact characteristics of these sequences are unclear because of the small number of such examples.]

The number and size of chromosomes have changed during mammalian evolution by rearrangement processes such as fusion, fission and inversion. Thus, the terminal portion (except for telomeres themselves) and internal portion have undoubtedly changed their positions during evolution. In spite of these global changes the segmental features of chromosomes have been stably preserved. For example, banding patterns in chromosome segments, as well as local gene organization and affiliation with the GC-rich or AT-rich genome compartment, have remained fairly constant during mammalian evolution (2,27). Certain GC-rich terminal portions of human chromosomes may thus possibly correspond to the internal T-bands of other mammals. Actually, very GC-rich internal bands (and thus T-bands) of the human genome were known to often correspond to terminal bands of other mammals and vice versa (11,15,23,28). This is the reason the vague expression, 'T-bands and/or terminal R-bands' is used in this paper. Systematic analyses of chromosomal locations of evidently GC-rich genes of various mammals may provide additional data on evolutionary processes at the chromosome level.

Biological meanings of compositional compartmentalization

Five types of isochores with different G+C% levels have been pointed out for the human genome (1,2). The GC-richest isochore, H3, is believed to correspond to T-bands, and gene concentration in the H3 is much higher (by a factor of 5–10) than that in other isochores (2). High gene density at particular genome portions indicates the regions to potentially be very active in transcription. Bloom syndrome mitotic chiasmata are prevalent in T-bands (11), indicating the possible importance of the regions also in recombination processes. Differences in G+C% levels among chromosomes (Table 2) may be due to differential populations of isochore classes in individual chromosomes but lack of a certain isochore class in some chromosomes appears still a possibility. Information by analyses on codon usages is inevitably biased to that on genome compartments carrying active genes. Information on compartments in which active genes are absent or very rare will undoubtedly be obtained by sequencing randomly cloned genome fragments as planned in the Human Genome Project.

The resolving power of banding techniques is limited to the level of Mb, and compositional maps with finer resolution can not be attained with them. It should be stressed that the purpose of studying long-range compositional distribution, based on nucleotide sequences, is not only to study their relation to cytogenetic data but also obtain more detailed information on compositional compartments (e.g. precisely locate compartment borders; 5,9) and clarify uncertainty of banding techniques. It

may be also possible to know the functional significance of compositional compartmentalization on a molecular basis (e.g. the functional relationship to distinct chromatin domains involved in replication and gene expression). Overall G+C% of the human genome is roughly 40%, and extreme GC-richness such as 80% or more at the third codon position should thus be due to clear factors. G+C% levels of introns and the genome portion flanking GC-rich genes are usually 55–65% and do not reach 80% (4,5). Thus, directional mutation pressure, even compartmentalized along the genome, can not explain the extreme characteristics at the third position, though directional mutation pressure may have acted during evolution (29,30). The biological significance of compositional compartments of various levels has been discussed by several authors (2,6,11,31–33). Thus, points less discussed are considered here. Not only metaphase chromosomes but also interphase chromosomes are known to be compartmentalized in the respective nuclei. Terminal positioning in chromosomes should be an important signal for nuclear location (34), and evident GC-richness may make the signal much clear. Internal T-bands may constitute analogous signals for nuclear positioning of long chromosomes. This is consistent with the fact that internal T-bands of human chromosomes often correspond to terminal bands of other mammals.

ACKNOWLEDGMENTS

The authors are very grateful to Drs. M.Kimura, S.Osawa, and G.Bernardi for encouragement and for helpful discussion. This work was supported by a grant-in-aid for scientific research from the Ministry of Education, Science, and Culture of Japan.

REFERENCES

1. Bernardi,G., Olofsson,B., Filipski,J., Zerial,M., Salinas,J., Cuny,G., Meunier-Rotival,M. and Rodier,F. (1985) Science 228, 953–958.
2. Bernardi,G. (1989) Ann. Rev. Genet. 23, 637–661.
3. Ikemura,T. (1985) Mol. Biol. Evol. 2, 13–34.
4. Aota,S. and Ikemura,T. (1986) Nucleic Acids Res. 14, 6345–6355, and 8702 (correction).
5. Ikemura T. and Aota S. (1988) J. Mol. Biol. 203, 1–13.
6. Mouchiroud,D., Fichant,G. and Bernardi,G. (1987) J. Mol. Evol. 26, 198–204.
7. Bulmer,M. (1987) Mol. Biol. Evol. 4, 395–405.
8. Filipski,J., Salinas,J. and Rodier,F. (1987). DNA 6, 109–118.
9. Ikemura,T., Wada,K. and Aota,S. (1990) Genomics 8, 207–216.
10. Gardiner,K., Aissani,B. and Bernardi,G. (1990) EMBO J. 9, 1853–1858.
11. Holmquist,G.P. (1990) In Obe,G. (ed) Advances in Mutagenesis Research vol 2. Spring-Verlag, Berlin Heidelberg, pp.95–126.
12. Dutrillaux,B. (1977) In Yunis,J.J.(ed) Molecular structure of human chromosomes. Academic Press, London New York, pp.233–265.
13. ISCN (1981) Cytogenet. Cell Genet. 31, 1–23.
14. Wada,K., Wada,Y., Doi,H., Ishibashi,F., Gojobori,T. and Ikemura,T.(1991) Nucleic Acids. Res. 19, Supplement, 1981–1986.
15. HGM10.5 (1990). Cytogenet. Cell Genet. 55, 1–778.
16. Kimura,M. (1983) The neutral theory of molecular evolution. Cambridge Univ. Press, Cambridge.
17. Sharp,P.M. and Li,W.-H. (1986) J. Mol. Evol. 24, 28–38.
18. Bulmer,M. (1988) J. Evol. Biol. 1, 15–26.
19. Yunis,J.J. (1981) Hum. Genet. 56, 293–298.
20. Comings,D.E. (1978) Ann. Rev. Genet. 12, 25–46.
21. Wada,K., Watanabe,I., Tsuchiya,R. and Ikemura,T. (1991) In Kimura,M. and Takahata,N. (ed.) New Aspects of the Genetics of Molecular Evolution. Japan Sci. Soc. Press, Tokyo/Springer-Verlag, Berlin, pp.195–210.
22. Kuhn,E.M. (1976) Chromosoma 57, 1–11.
23. Ambros,P.F. and Sumner,A.T. (1987) Cytogenet. Cell Genet. 44, 223–228.
24. Paris Conference (1972) Cytogenet. Cell Genet. 11, 313–362.
25. Willis,A.E., Weksberg,R., Tomlinson,S., Lindahl,T. (1987) Proc. Natl. Acad. Sci. USA 84, 8016–8020.
26. Satchwell,S.C., Drew,H.R. and Travers,A.A. (1986) J. Mol. Biol. 191, 659–675.
27. Sawyer,J.R. and Hozier,J.C. (1986) Science 232, 1632–1635.
28. Dutrillaux,B. (1979) Hum. Genet. 48, 251–314.
29. Sueoka,N. (1988) Proc. Natl. Acad. Sci. USA 85, 2653–2657.
30. Wolfe,K.H., Sharp,P.M. and Li,W.-H. (1988) Nature 337, 283–285.
31. Holmquist,G.P. (1989) J. Mol. Evol. 28, 469–486.
32. Filipski,J. (1990) In Obe,G. (ed) Advances in Mutagenesis Research vol 2. Spring-Verlag, Berlin Heidelberg, pp.1–54.
33. Herbomel,P. (1990) The New Biologist 2, 937–945.
34. Cremer,T., Baumann,H., Luedtke,E.-K., Sperling,K., Teuber,V. and Zorn,C. (1982) Hum. Genet. 60, 46–56.